Globus.org: Hosted Service for Reliable Data Movement

Raj Kettimuthu

Argonne National Laboratory and

The University of Chicago

the globus alliance
www.globus.org

Today's Science Environments

- Large-scale collaborative science is becoming increasingly common

Fusion community's International ITER project

- Distributed community of users to access and analyze large amounts of data

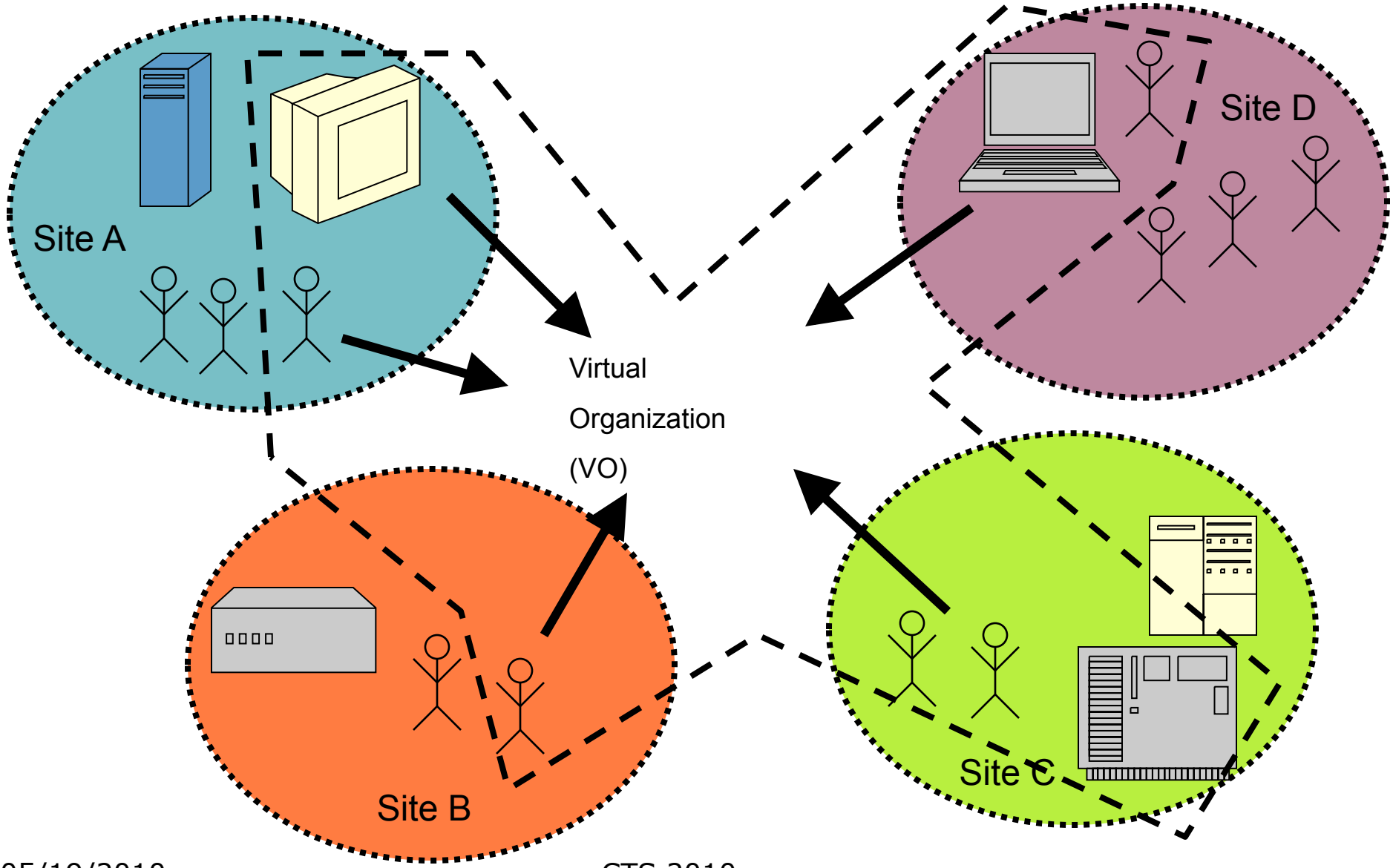
the globus alliance
www.globus.org

Large Hadron Collider

1800 Physicists, 150 Institutes, 32 Countries

100 PB of data by 20XX; 50,000 CPUs?

Virtual Organization



the globus alliance
www.globus.org

Data Movement

- Raw simulation or observational data is just a starting point for most investigations
- Understanding comes from further analysis, reduction, visualization, and exploration

Petascale resource

Compute Cluster

Scientist's Desktop

- Furthermore the data is a community asset that must be accessible to any member of a distributed collaboration

the globus alliance
www.globus.org

GridFTP

- High-performance, reliable data transfer protocol optimized for high-bandwidth wide-area networks
- Based on FTP protocol - defines extensions for high-performance operation and security
- Standardized through Open Grid Forum (OGF)
- GridFTP is the OGF recommended data movement protocol

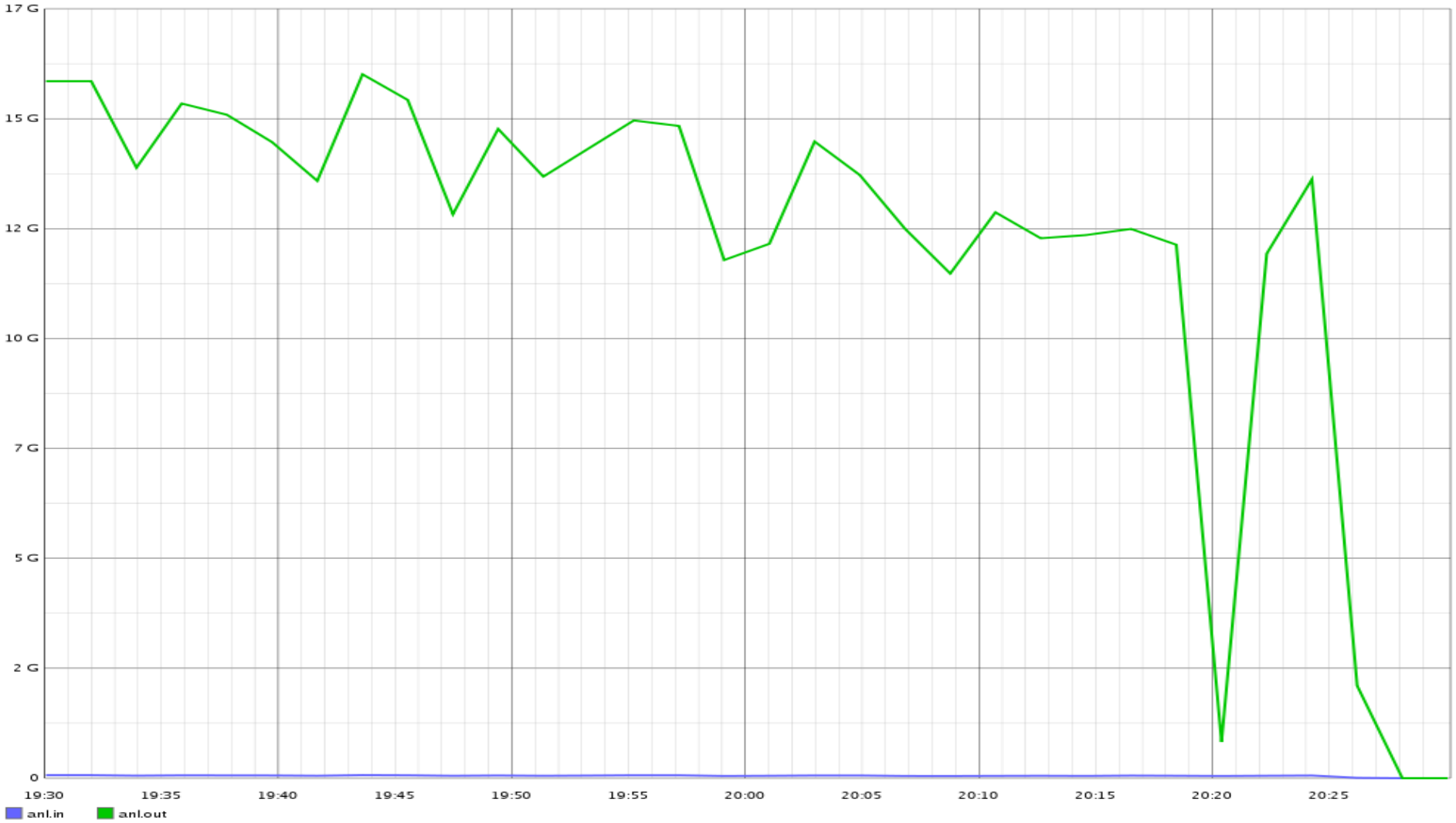
the globus alliance
www.globus.org

Globus GridFTP

- Performance
  - Parallel TCP streams, optimal TCP buffer
  - Non TCP protocol such as UDT
- Multi-node data movement
- Multiple security options
  - Anonymous, password, SSH, GSI
- Support for reliable and restartable transfers

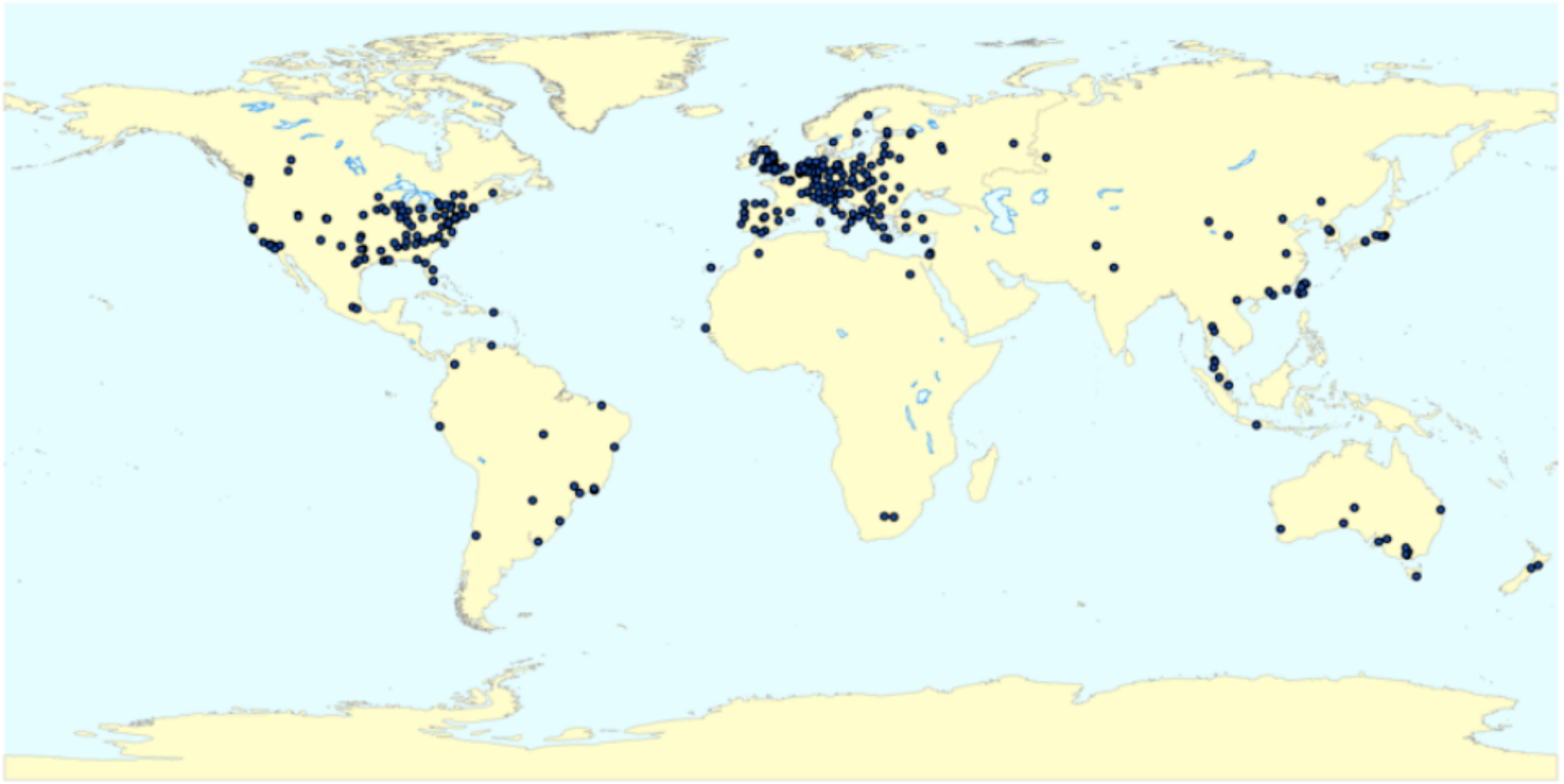
the globus alliance
www.globus.org

Performance



the globus alliance
www.globus.org

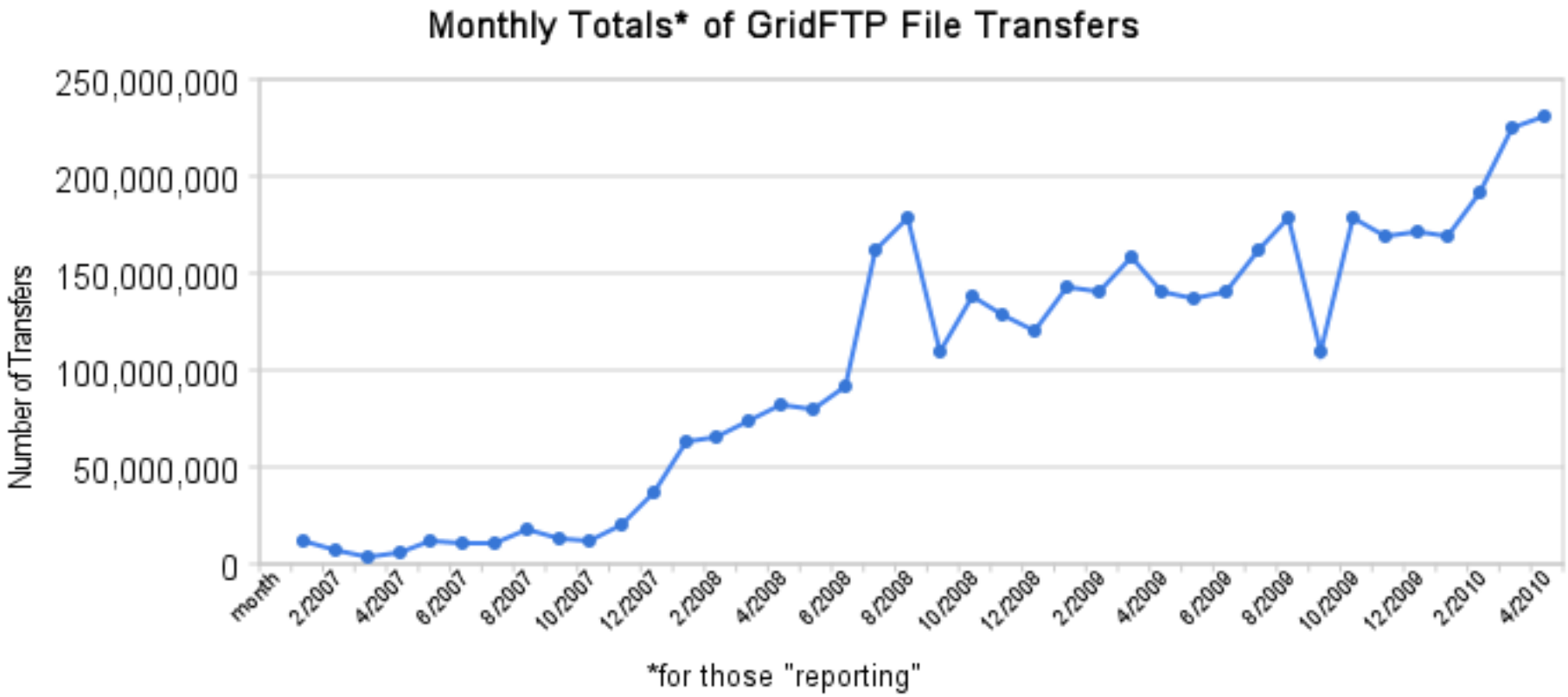
GridFTP Servers Around the World



Created by Tim Pinkawa (Northern Illinois University) using MaxMind's GeoIP technology (http://www.maxmind.com/app/ip-locate).

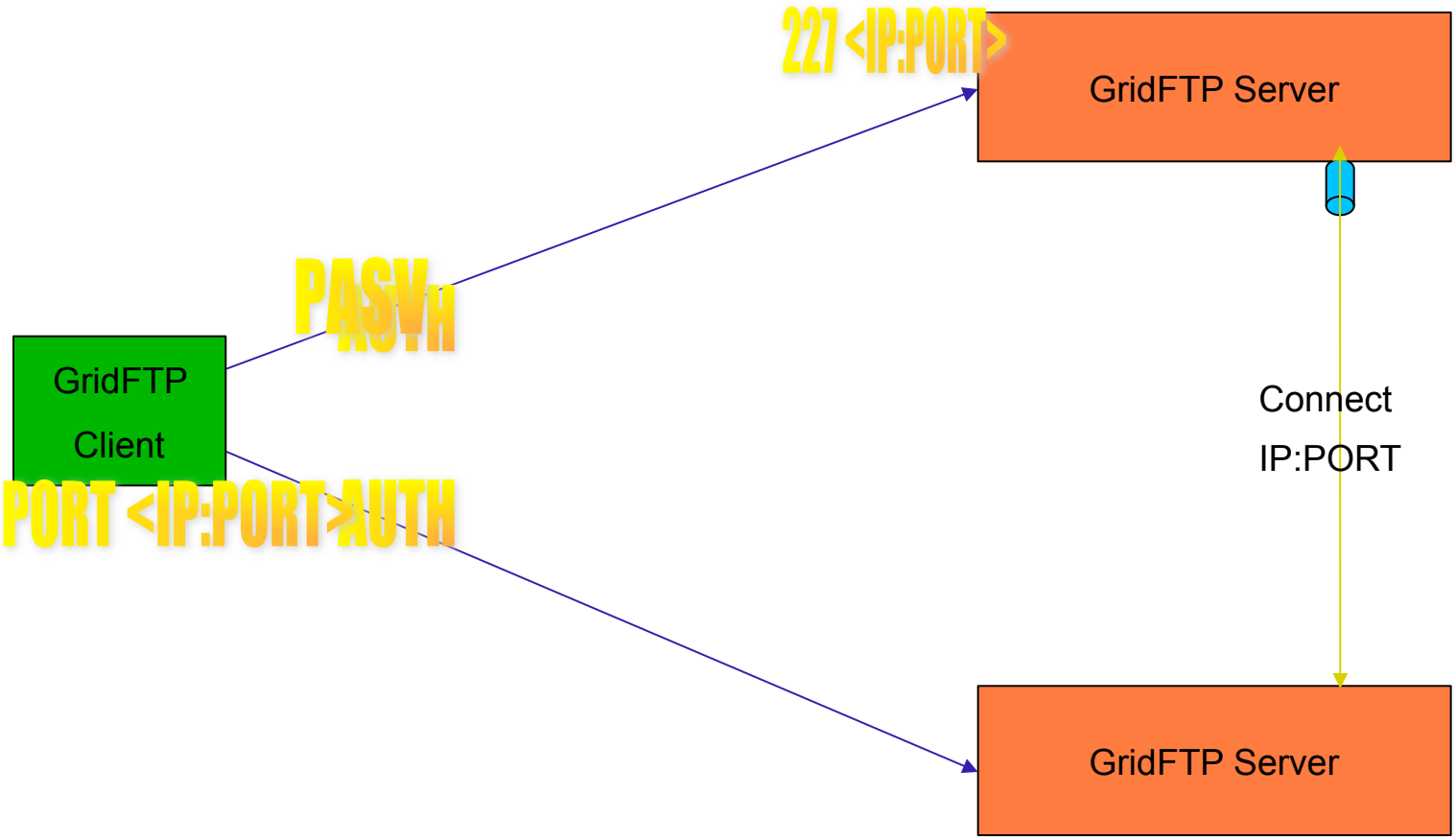
the globus alliance
www.globus.org

GridFTP Usage



the globus alliance
www.globus.org

GridFTP Data Movement



Globus.org hosted services: Data replication as an example

"Move data from A to B"

Globus.Org

Notification agents

Transfer agents

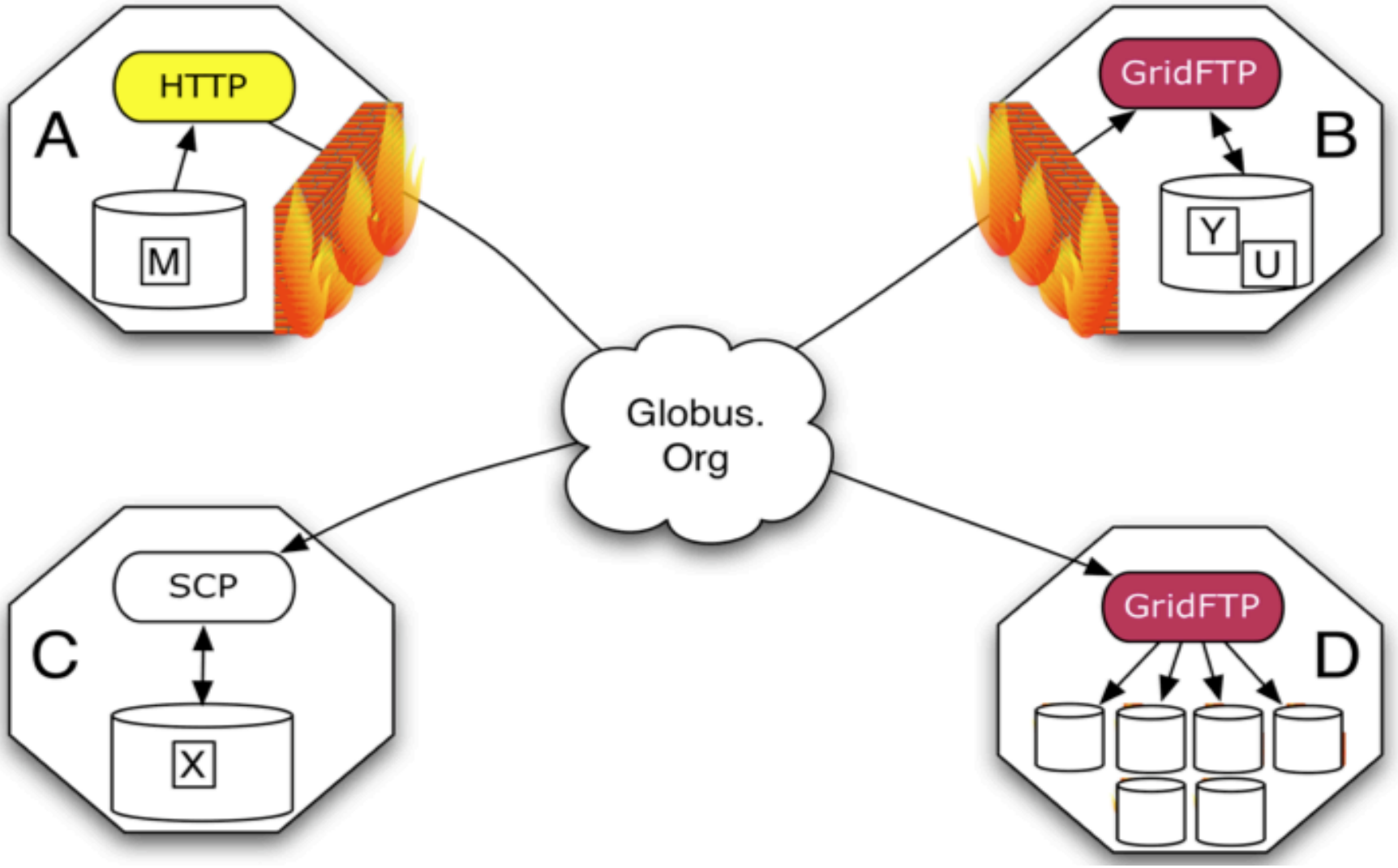
Persistent database

Amazon Web Services

A

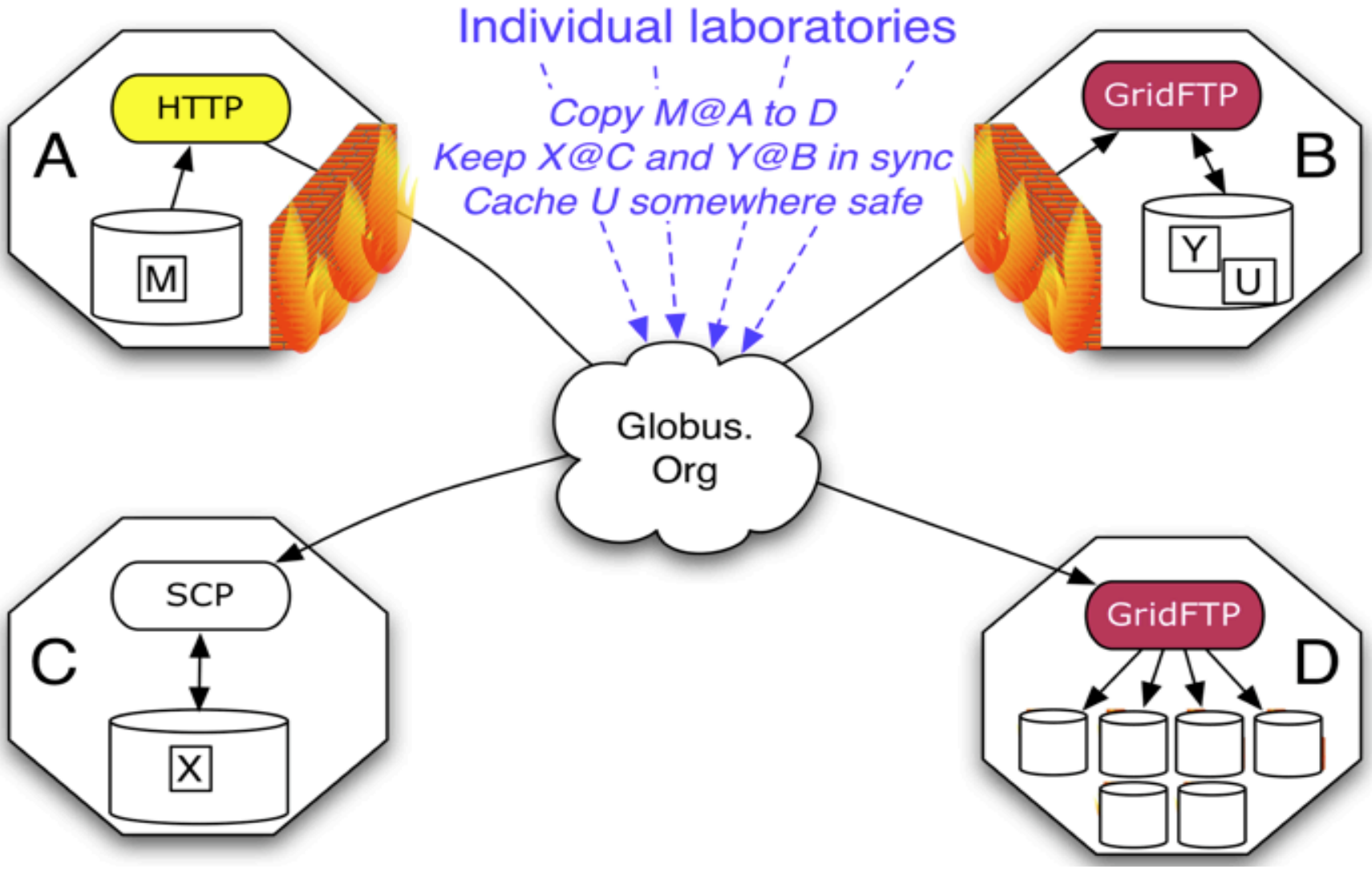
B

Globus.org



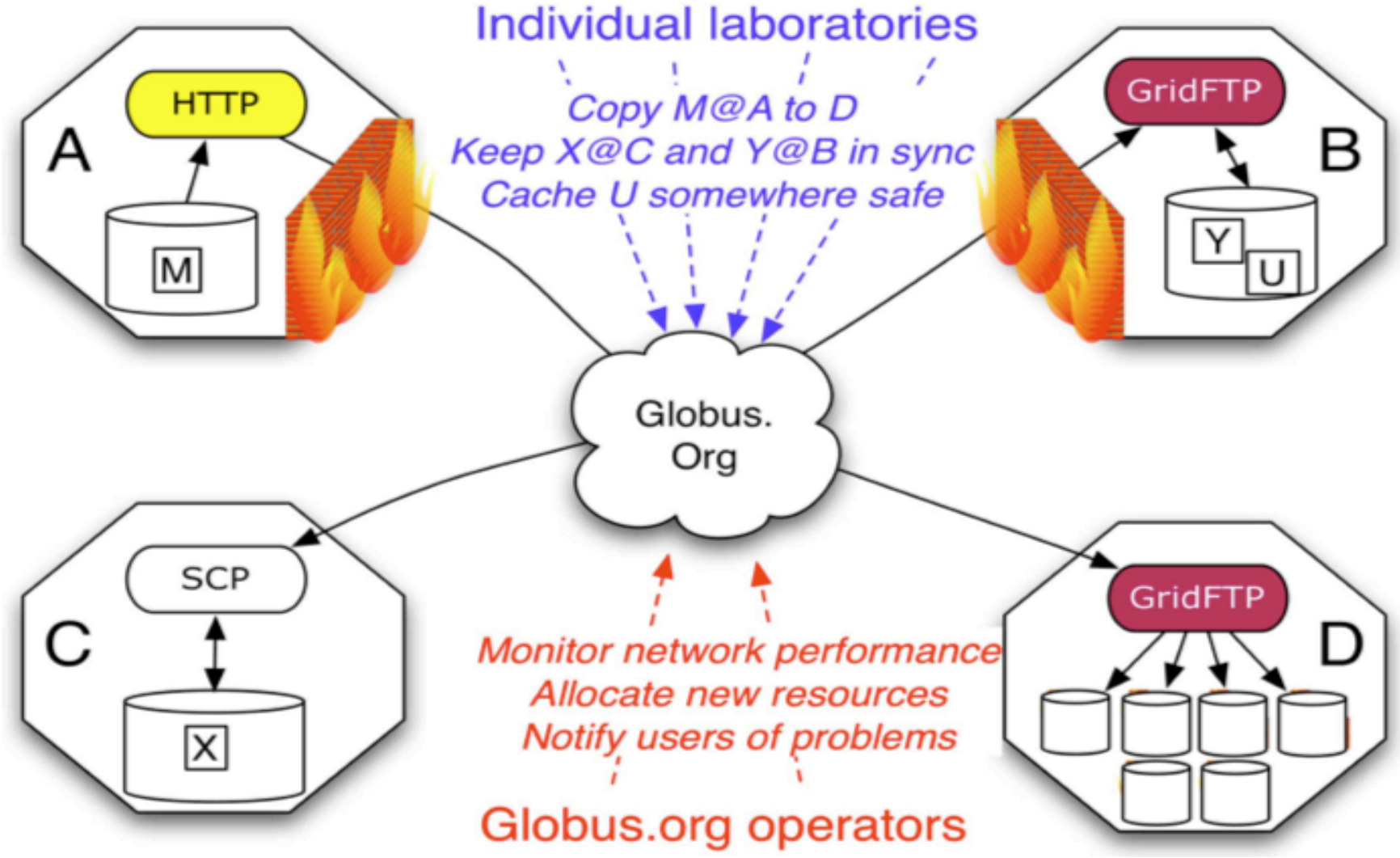
the globus alliance
www.globus.org

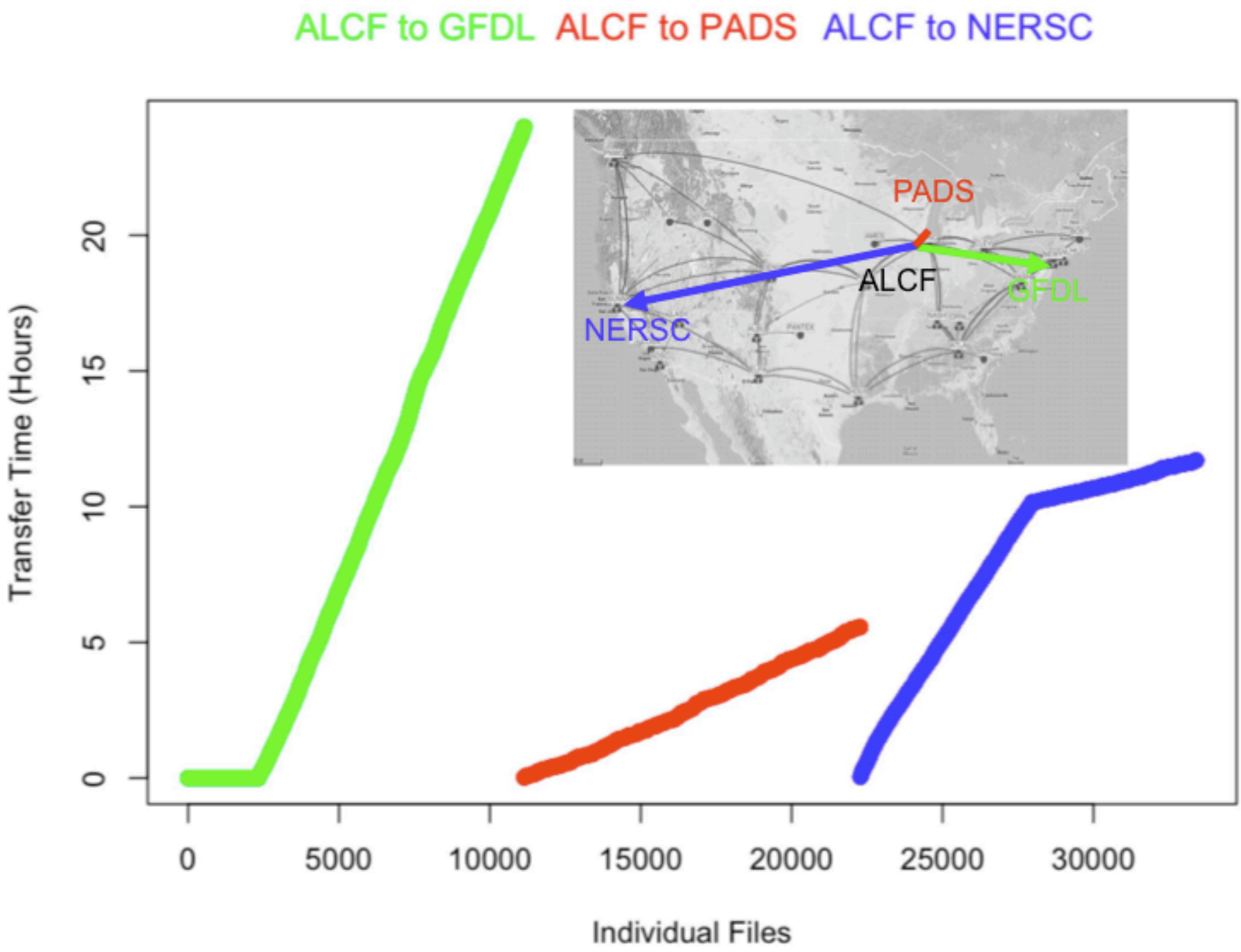
Globus.org

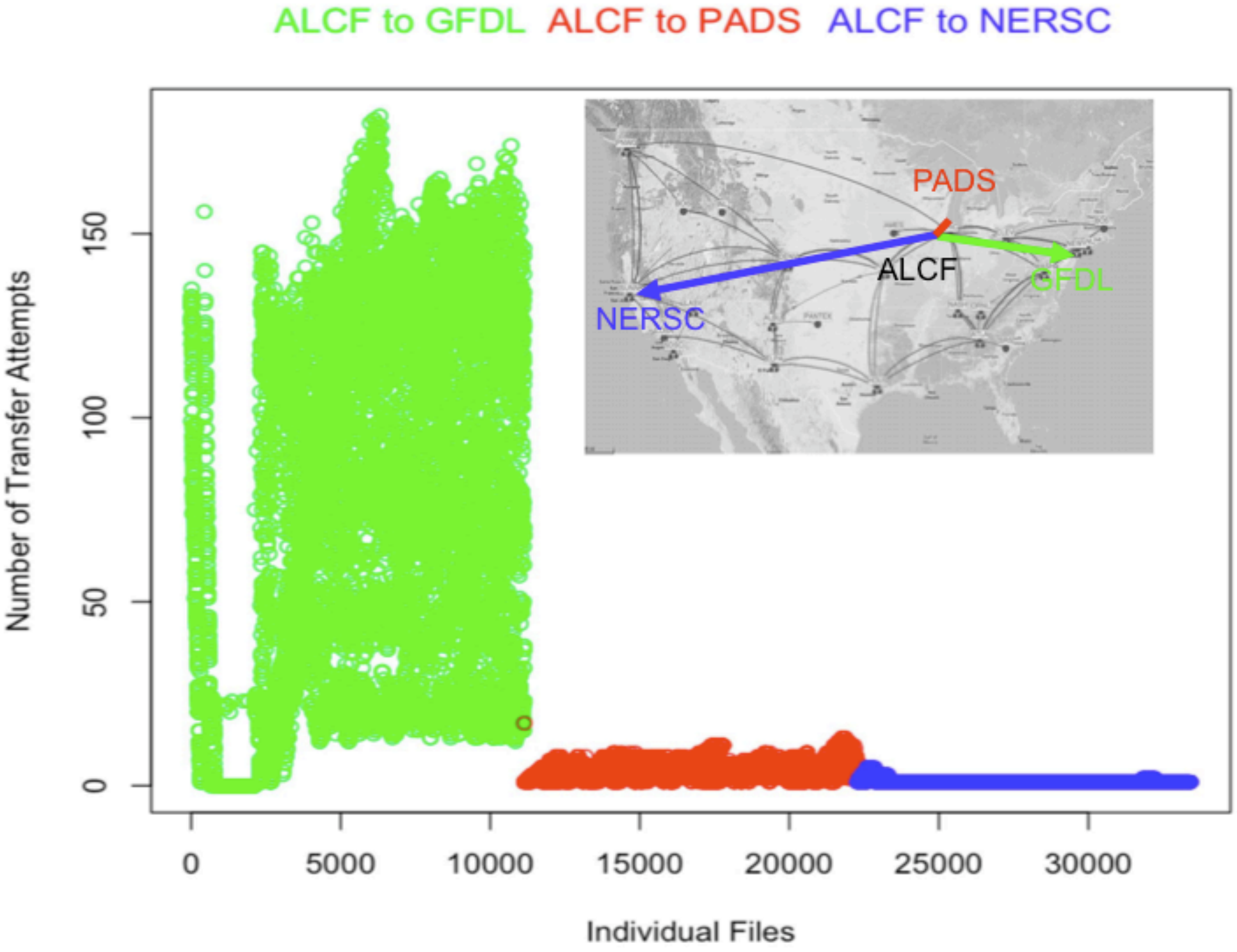


the globus alliance
www.globus.org

Globus.org




ALCF to GFDL
ALCF to PADS
ALCF to NERSC
PADS
ALCF
GFDL
NERSC
Transfer Time (Hours)
0
5
10
15
20
0
5000
10000
15000
20000
25000
30000
Individual Files


ALCF to GFDL
ALCF to PADS
ALCF to NERSC
Number of Transfer Attempts
0
50
100
150
0
5000
10000
15000
20000
25000
30000
Individual Files
PADS
ALCF
GFDL
NERSC

the globus alliance
www.globus.org

More Information at
http://www.globus.org/service/
http://www.gridftp.org

Questions